LBL-10423 c.2

Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

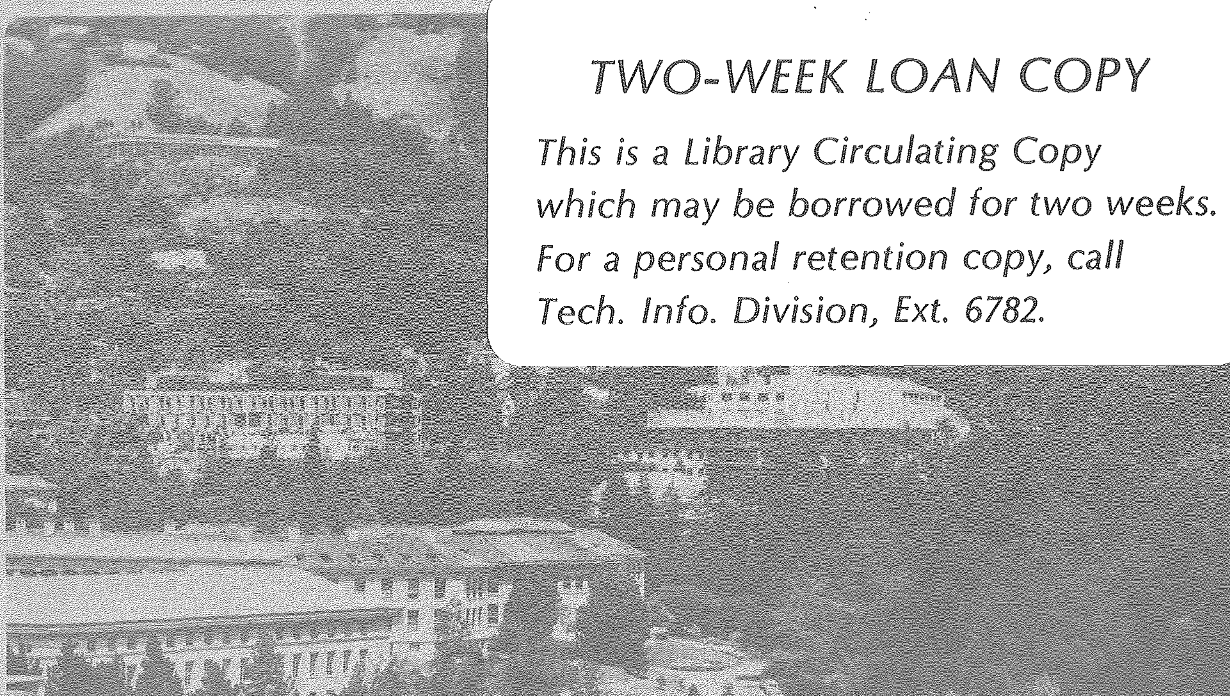
Employee & Information Services Division

Presented at the Nuclear Records Management Association National Meeting, San Francisco, CA, September 11-12, 1979

INDEXING WORKSHOP: HOW TO ASSIGN KEYWORDS

Virginia Sternberg

September 1979



TWO-WEEK LOAN COPY

This is a Library Circulating Copy which may be borrowed for two weeks. For a personal retention copy, call Tech. Info. Division, Ext. 6782.

LBL 10423 c.2

Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Presented at the Nuclear Records
Management Association National
Meeting, San Francisco, CA
September 11-12, 1979

LBL-10423

INDEXING WORKSHOP:

HOW TO ASSIGN KEYWORDS

Virginia Sternberg

September 1979

Prepared for the U. S. Department of Energy under Contract W-7405-ENG-48

INDEXING WORKSHOP:

HOW TO ASSIGN KEYWORDS

You have heard about issues surrounding indexing and retrieval of nuclear records and automation and micrographics of these records. Now we are going to get each of you involved in indexing and assigning keywords.

The first part of this hands-on workshop will be a very basic, elementary step-by-step introduction, concentrating on how to assign keywords. It is a workshop for beginners, People who have never done it before. It is planned to demonstrate what an analyst has to do to index and assign keywords to a document. Then I will take some pages of a report and demonstrate how I choose keywords for it. Then each of you will have a chance to do the same thing with similar pages from another report. Then we will discuss the variations in the keywords you individually assigned.

There are many systems that can be used. In this particular workshop we will cover only a system of building your own keyword listing as you index your documents.

We will be discussing keywords or descriptors or subject words, but first I want to point out a few other critical points about indexing.

When developing an indexing project the most important thing to do first is decide what elements you want to retrieve by. Whether you go into a large computer retrieval system or a small three-by-five card system, you have to decide in advance what you want to retrieve. Then you can go on from there.

If you only need to search by equipment number or by purchase order or by contract number, then you can use a very simple retrieval system. But if you want to be able to retrieve a record by any combination of elements, then you have to consistently input these into your system. For example, if you want to be able to ask for the drawings of the piping in the secondary cooling system, level 3, manufactured by a certain vendor, then you must have put the information into the index by a retrieval file point, in advance.

I want to stress that the time spent in deciding what has to be retrievable is never wasted. It will save going back later to re-do thousands of records.

Let me quickly read a list of some of the elements to index: Personal author, corporate author (including contractor, supplier, manufacturer or designer), contract number, report number, purchase order number, equipment number, part number, revision number, building number, room number, level number, drawing number, specification or standard number, date, change notice number, and even date of change notice.

There are probably many other elements you have identified for your own records. To be able to retrieve by any combination of these elements is invaluable. Then, if you add subject retrievability to these elements, you have a very comprehensive system.

The elements mentioned are usually readily discernable by skilled non-technical people who are familiar with the paperwork issued <u>by</u> and flowing <u>through</u> their organization. Therefore, no matter how many of these elements are determined to be necessary, they can readily be identified and incorporated into the indexing form or coding sheet.

Now back to keywords.

PURPOSES OF KEYWORDS

The keywords or descriptors elucidate the ideas in a record and are clues that lead someone else to it in the future. This may sound like a very simple task, but it isn't.

The use of keywords in retrieval systems has increased in the last 25 years. The concept of indexing by keywords and then combining the keywords to find the few documents that deal with all the subjects desired has been used extensively for indexing technical literature. The success of these technical indexes has led information specialists in other subject areas to start similar keyword systems. And as you see from this meeting, records managers too. The various reasons for assigning keywords are:

- To provide indexes for searching
- To retrieve by subject
- To retrieve quickly
- To retrieve by using terms familiar to the searcher.

A further explanation of each of these purposes seems appropriate at this point.

To Provide Indexes

The key purpose in assigning keywords is to produce an index which a searcher can use in the future to find information he needs. This obvious statement may seem superfluous but it must be kept in mind, because there are times when analysts forget that the purpose is to help the searcher of the future, not just assign keywords for a system.

The person assigning the keywords must be the liaison between the creator (the author, the source of the document) and the searcher or receiver of the information in the document. The assignment of keywords is a great responsibility. The analyst must view the document in hand as a communicator of potential information which will assist someone in the future.

To Retrieve by Subject Content

Another primary purpose in assigning keywords is to describe the subject content of a document in such a manner as to facilitate its retrieval efficiently and without fail in the future. Keywords have to be assigned in an effort to expedite retrieval of the subject matter regardless of the terminology the searcher uses.

Thus, the keywords must (1) cover the specific subject as well as the general subject, and (2) be highly technical for the expert and semi-technical for the layman.

The purpose in assigning keywords is to provide a subject description of documents so they can be located in the future. The determination of the main points in a document is the secret of any retrieval system. The assignment of the most appropriate descriptor or keyword can lead a searcher in the future to a meaningful report, drawing, or piece of correspondence.

The idea in assigning keywords is to increase the probability of delivering the needed information to the inquirer whenever he searches the data base.

To Retrieve Quickly

Another purpose in assigning keywords is to provide a quick and easy searching device for the user. A good index, which is easy to use, helps a user when he is searching for a document or a drawing. Most indexes in the past have been difficult and time-consuming to use because they were contrived to be hierarchical or decimal or they were too general.

To Retrieve by Using Terms Familiar to the Searcher

Another purpose for assigning keywords is to provide a search tool which uses terms that are commonly used by those working in the field. Keywords are assigned using terms familiar to these people. They are terms taken right from the documents being indexed and subsequently to be retrieved. When this type of terminology is used for an index or retrieval system, a search can often be made without consulting any kind of dictionary (or thesaurus).

This use of normal terms for keywords encourages a searcher to return again to search for other subjects. Nothing discourages a searcher more than finding his subject indexed under an all-encompassing general subject or some term that is obsolete or some arbitrary term taken from a thesaurus. We are developing a system for the user and we want him to come back.

Now that we have established the purpose of assigning keywords I want to explain the variations in keywords and explain exactly what is meant by the term keyword.

A keyword is a word or phrase which describes the essence of the subject content of a document and will permit retrieval by that subject. It describes a very specific object or action. In combination with other keywords, it will direct a person to a specific document. The ideal keyword is the one which communicates the information to the potential user. A keyword can be considered a subject word which is used to bring like documents together in a retrieval system.

The following classes can be kept in mind when establishing a new keyword to use for identifying a new and relevant subject. There is no significance to the term "Class" as I use it here. I believe, however, that keywords fall into groups such as the following:

1. Machines, apparatus, devices, tools, buildings
2. Processes, reactions, operations, events, procedures, tests
3. Materials, substances, chemical products, material elements, liquids
4. Attributes, characteristics, qualities, measurements, effects
5. Place, time, conditions
6. Abstract concepts

I would like to comment on each of these six classes.

1. <u>Machines, Apparatus, Devices, Tools, Buildings</u>

The most common keywords are terms describing physical objects. A person can touch or see a valve, a cross-beam, a dump-heat-exchanger, a gamma-shield.

2. <u>Processes, Reactions, Operations, Events, Procedures, Tests</u>

A process such as decay-heat-removal can be observed as it is happening. A recovery operation is an operation that can be described and has steps that can be seen.

3. <u>Materials, Substances, Chemical Products, Material Elements, Liquids</u>

Rain is a physical, tangible liquid that can be seen and touched; it is <u>concrete</u>. Water, steam, sleet, snow, ice, and hail are related terms that can also be used for keywords. If a report discusses what effect rain has on a building, a specific keyword to assign would be "rain," not "water."

4. <u>Attributes, Characteristics, Qualities, Measurements, Effects</u>

Design-life, design-description, system-configuration, wind-loads are examples of terms that describe a measurement or a quality.

5. Place, Time, Conditions

Searchers ask for temperature requirements, date, or a site or state. Therefore, keywords specifying such items are helpful in retrieving the right document. Examples of conditions could be acceptance-criteria, environmental conditions, operating conditions, performance-standards, shipment problems.

6. Abstract Concepts

I pointed out earlier that a keyword should be a term that describes something tangible. So, why should "abstract concepts" be included? There are always exceptions to every rule. There are theories, ideas, and concepts that can be used as keywords. For example, Einstein's Theory of Relativity can be indexed by using the keywords Relativity Theory or Theory of Relativity. However, it is advisable to use a hardware term or a process, or a machine, whenever possible.

This discussion of classes of keywords leads up to a more detailed description of keywords.

Keywords will lead a searcher to the document. The document in turn will potentially give him the information he is looking for. So keywords should be assigned the way searchers think when they ask a question. Keywords should be terms used every day by those working and writing in the specific field. The terms that are familiar to a worker in the field of health safety are the keywords he will use when he asks a question.

Keywords can also be terms that are unique to an organization. They can be trade names or words describing equipment, even patentable items. Keywords are terms which describe something tangible, as I mentioned before:

- A component
- A piece of equipment
- A process
- A room
- A building or
- A piece of metal

Keywords are usually nouns, for example:

- Generators
- Hangers
- Instruments
- Pipes
- Pumps

Keywords are usually plural words. When a searcher asks a question, he may ask for all the drawings on Valves, or Gate Valves, or Isolation Valves, or Pumps, or Centrifugal Pumps. Some keywords are words or phrases that are always used together. Some examples are:

- Heat-transfer
- Liquid-metal
- Heat-exchangers
- Leak-detectors

Now, with the explanation so far, do you think you can assign keywords?

Basic Rules to Follow

There are four fundamental rules which must be adhered to in assigning keywords. These four rules take overriding precedence.

1. First of all be specific. Keywords are assigned on the exact subject of the document. For example: "check valves."
2. Second, assign general terms. For example: "Valves" is an inclusive term. If a searcher wants everything on valves, all kinds of valves, this term would be the keyword that he would use.
3. Third, be consistent. Use the same term for the same component every time. The "name of the game" is consistency. In assigning keywords, analysts (a) must be consistent, (b) be consistent in assigning generic terms, and (c) be consistent with each other.
4. Fourth, be as complete as possible without going to extremes.

In addition to the four fundamental rules there are many more specific, detailed rules to follow. Some of these are:

- Use common words regularly used by the writers.
- Use words that are not ambiguous.
- Use scientific names.
- Use engineering terms normally used by engineers.
- Use current acceptable terms, not obsolete terms.
- When a subject is covered by several terms, choose one and record it in the thesaurus. List the rejected terms and refer users to the term decided upon.
- Assign generic terms, that is, more general terms, or more inclusive terms, even though not in the document.
- Do not use terms which are lower on the hierarchical scale than the information in the report. Choose keywords only for information in the document although it has been suggested above that generic terms

be used; but it is not worthwhile to add keywords for every nut and bolt in a component unless they are actually discussed in the document.

- Add a modifier in parentheses after a term when it has two meanings, to denote the specific application. For example, Beams (Structural) or Beams (Photon).
- Use related words that define a term if these words are regularly used by the people who will use the system.
- Avoid the use of vague words such as studies or reports.

I think I can hear you saying, "Now I can assign keywords."

STEPS IN ASSIGNING KEYWORDS

Each one of the steps will become automatic within a short period of time for most analysts. None of these steps has a special formula or has to be learned. And it is helpful to have an education in the subject field or have experience working in it.

1. Evaluate Document

The first step in assigning keywords must be to ask the question, does this type of document go into the system. If so, is it worth putting this particular one into the system?

The most important decision to make when a document is being reviewed for the assignment of keywords is whether the document is worth spending the time to do it. Every document had a purpose when originated. However, there may be little need for it in the future. A classic example is the agenda for a meeting. Usually minutes of the meeting are issued later. There is little point in recovering the agenda at some future date. The minutes record what actually took place and are valuable when questions are asked about when and why a decision was made.

A list of the types of documents to be entered into the system and assigned keywords should be compiled and made available to the analysts. The form of a document, e.g., film, fiche, hard copy, videotape, drawing, or letter, should be irrelevant. For the user there is only one criterion--does it contain information. The decisions of what documents should go into the system should be made by well-informed people who are knowledgeable on the subject in each type of document and the overall mission of the retrieval system. This list should be prepared in advance.

It takes a considerable amount of time to review a document, assign keywords, and keypunch the information or prepare it for input into a system. Therefore, some judgment should be made concerning each document prior to input.

In many cases it may be faster and less expensive not to make an evaluation of every document. In some cases even though an overall preliminary decision has been made to index certain categories there may be times when one of these contains little information. The analyst should be encouraged to make the decision <u>not</u> to spend the time to work on those documents.

When a decision has been made to <u>not</u> assign keywords, a procedure needs to be established for handling the documents. Should they be entered into the system at all? Should a minimal amount of information be entered? Or should they be eliminated entirely? Such a procedure should be prepared in advance.

2. <u>Determine Descriptors for Document</u>

A detailed listing of fields or elements for which to prepare input should also be provided for the analysts. Again, the decision as to what types of information are to be retrievable should be made by those who know the subject matter <u>and</u> the project <u>and</u> the future user.

Based on these lists of fields, the analysts should take each document and review it for overall descriptive information, determine the various items required and record them on the input form provided, or identify them on the document itself. Examples of elements for which a field has been designated can include items such as:

- Corporate Author
- Personal Author
- Drawing Number
- Record Type
- Revision Number or
- Date

These items could also include any other additional information which would assist the searcher to locate this document in lieu of or in parallel with keywords.

3. <u>Review Specific Parts of Document</u>

The analyst should search for and review the following sections of each document in order to quickly find the significant points being made by the author:

- Title
- Introduction

- Preface
- Figures
- Tables
- Section of Results
- Conclusions
- Recommendations
- Appendices
- Attachments

The reasons for reviewing these parts of a document are:

- To obtain an idea of what the author intended to put into the document.
- To find a short summary of it.
- To find out the purpose of the work.
- To get a birds-eye view of the final result of the work done, or to be done, or redone, or to be changed, cancelled or initiated.

5. To identify a significant number of keywords.

4. <u>Read the Pertinent Sections of Document</u>

The following are "musts":

- Abstract
- Recommendations
- Conclusions
- Table of Contents
- Title Block (if a drawing)

Perusal of these major sections provides the analyst with a general overall summary of the information in the document. After this review he can determine if he understands the subject matter at all or if he has to consult a dictionary or handbook just to understand the terminology.

5. Assign Preliminary Keywords

The analyst should jot down words as he reads the abstract, recommendations and conclusions:

- All the possible pertinent subject words as they come to mind.
- All the keywords from the abstract, conclusions, etc.
- All obvious keywords.
- Any special points being emphasized.

Don't think about the form of the words--just write them down. The exact form of the keywords will be considered later. The goal at this time is to find all of the key pieces of information which should be retrievable in the future.

6. Study Document for More Information

Study the document more carefully:

- Ask questions such as: What is this document really about? What question is this piece of paper going to answer? What is important in this drawing? What in this document may be important to retrieve in the future?
- Ask more detailed questions: Is this in the Primary System? Is this in the Loop? Where is this cold trap? How does this freeze vent really work? What is a cold-leg-isolation valve?

Then, if, after studying the document carefully, a certain part is not clear or is not understood, consult the author. Only if in doubt call the author. This is not to be resorted to automatically. When consulting him, point out (1) what you are doing, (2) why you are doing it, (3) the document being indexed, (4) the keywords already considered, (5) the specific question you need to have answered.

7. Assign Additional Keywords

Write down the additional keywords resulting from a more intensive study of the document, discussion with the author, and any discussions with other analysts. These are off-the-cuff, analytical ideas about the subject content of the document. The object is not to remember the form of the keyword but to quickly gain the maximum number of important subjects in the document.

These "quick and dirty" assignments of keywords eliminate the review and re-review of a document which would be the case if every keyword were checked in the thesaurus as the subject is determined.

It has been demonstrated that as analysts gain experience, they automatically assign keywords in proper format. This comes with time and should not be the first step in assigning keywords. The subject itself is king. Analysts should determine that first, and revise keywords later.

8. Make Final Selection of Keywords for Document

Revise the list of keywords.

- Make sure they are consistent; e.g., are they plural? are they nouns?
- Check them against the thesaurus of terms or the dictionary of words used.
- Determine where or how the keyword was used before.
- Make a judgment about adding the general keywords.
- Make a decision as to which specific modifying terms should be added to the keywords if any.

This is the point where the consistency is finalized, where the keyword assignment makes or breaks a system. The analyst must ask the question, "Are these the terms that a searcher would use if he were looking for this specific document?"

9. Check again: Is That all There is?

Now ask again, is that all there is in this document? Review the abstract, conclusions, and recommendations. If a drawing, review the title block. If a letter, review the introductory paragraph. If a specification, review the purpose.

Just one more time, the analyst should review the keywords and ask himself the question, "Is that all there is that is significant in this document?"

Emphasis of this point is not being made to suggest that large numbers of keywords be assigned to every document. Each document should have the number of keywords it deserves. A two-page letter of historical significance may require more keywords than the minutes of a lengthy meeting where only one decision was made.

Analysts should constantly remind themselves that assigning keywords is not a game of numbers. They must remember that leading a person to hundreds of documents with a few sentences on a subject is not a good service.

COMPILATION OF A THESAURUS OR DICTIONARY

The purpose of a thesaurus or dictionary of keywords is to assist the analysts in being consistent and to verify words when doing a search. However, you must be sure that the keywords in the thesaurus are not the only ones considered. Analysts must be aware of the value of adding to the thesaurus when a new word seems appropriate.

When you are considering starting an information retrieval system, look for a thesaurus that covers the subject matter of the documents. If a detailed thesaurus is not located it is advisable to start one rather than use an inadequate one.

Most dictionaries or thesauri of keywords already in print should not be depended upon, however, because their terms are so general. For example, the Engineering Thesaurus does not list the keyword Brackets; and therefore not the more specific keyword Support arm brackets. In order to find documents about Support Structures, the searcher would have to look at the titles of all documents listed under supports or mechanical structures.

In starting a new thesaurus I recommend that it be developed starting with the first document and the assignment of the first keyword. For each new document, keywords are assigned and added to the thesaurus. Consistencies and inconsistencies will be recognized easily as it is compiled.

The analysts should know that they are developing a new dictionary that is alive and expanding on a daily basis because it is a list of words being used by those working in the field, those writing the documents and probably searching for them in the future.

Successful retrieval systems do not restrict the keywords or the vocabulary in the dictionary. Rules are developed (1) to encourage the growth of the vocabulary and (2) to provide cross references for synonyms.

Standard practice is to list the new keywords on cards and interfile them daily. This is done to provide the analysts with a continually up-to-date thesaurus. If this keyword file is a punched-card file, a printout can be run off weekly. The ideal situation is to have a thesaurus on-line to a computer where it would be always up to date, new keywords can be added anytime, and users can also access it.

This concludes the basic introduction. Now I will demonstrate assigning keywords to a report. Then you will have your chance.